Turing's Test and Believable AI in Games

DANIEL LIVINGSTONE
University of Paisley, Paisley, UK

The Turing test is perhaps the most famous, most quoted, and probably most often misrepresented and misunderstood measure of machine intelligence. In this article we'll briefly review the Turing test and some of its key criticisms. In particular, we will try to answer whether the Turing test – or something derived from it – can be of use in developing and assessing game AI. We will also present a brief overview of a methodology for conducting believability testing for games and highlight some of the problems inherent in any attempt to categorically determine whether or not some AI behavior is capable of convincing, life-like behavior.

Categories and Subject Descriptors: I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *Games*; J. 7 [**Computer Applications**]: Computers in Other Systems – *Consumer products.*

General Terms: Design, Human Factors, Measurement

Additional Key Words and Phrases: Games, believability, evaluation

1. THE TURING TEST

The Turing test usually takes the form where an interrogator in one room uses a computer terminal to play a game of question and answer with two subjects who are located in another room. One of the subjects is human while the other is a machine; the task of the interrogator is to determine which is which. If the interrogator is unable to tell, then the machine must be considered intelligent.

Indeed, a Turing test, much like that described above, occurs annually, with a number of judges conversing in type via computer with talking computer programs and human confederates. To date, no program entered for the Loebner prize[1] has managed to fool the judges into thinking that it is human – and the prospects of such happening in the near future are arguably remote. This despite the date by which Turing expected his challenge to be met:

> *"I believe that in about fifty years it will be possible to program computers… to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning." [Turing 1950]*

It appears that meeting the challenge set by Turing is significantly harder than he expected. Recent *winning* programs in the annual Loebner prize competition scored in the range from *probably* a computer program to *definitely* a computer program (while the human confederates in the 2003 competition scored in the range *undecided* to *probably human* – which may itself tell us something about how the Turing test limits our ability to recognize intelligence when we do encounter it [2]). Indeed, even if the challenge had been met, strong arguments have been made that we still would not necessarily have built an

Author's address: University of Paisley, Paisley, United Kingdom

intelligent machine (the most famous refutation being Searle's [1980] "Chinese Room Argument.") The basis of Searle's argument is that a program could pass the test – outputting strings of symbols for each string of symbols input – without any understanding of what it was doing, and hence pass without possessing intelligence.

While the test described above does capture Turing's (apparent) intent, it differs significantly from the Imitation Game that Turing seems to propose. The Imitation Game is to be played by an interrogator, a man, and a computer program. The man and the computer program compete as to which one of them is more convincing as a woman. Further, in the Loebner test, all human judges know that they are interrogating a mix of humans and machines; but in Turing's game it is not even clear whether the judge is aware of the possibility that one contestant is a machine. Turing does not elaborate on the subtle conversational gambits required for this gender test. The inaccurate version of the test in which the computer has simply (!) to out-human the human has become the conventional one, and is now considered by most to be *the* Turing test. The significance of the differences between the conventional Turing test and the gender test is discussed by Hayes and Ford [1995], who argue against the use of the Turing test in its popular form as a sensible goal for AI research.

Yet the Turing test maintains its appeal, at least in the popular press if not among AI researchers, who are increasingly turning to other so-called "grand challenges." These new challenges focus more on behaviors embodied in the real world, and, unlike the Turing test where machines either pass or fail, consist not only of a long-term objective that must be met but also provide many incremental goals, allowing progress to be made in a systematic and directed manner [Cohen 2004].

Like Turing's test, however, these challenges still focus on the ability of machines to complete tasks that appear to demonstrate, not prove, intelligence. Turing was able to see that formal definitions of intelligence were likely to remain troublesome, and so it might never be possible to formally *prove* that a machine is intelligent. Instead, we should try a different problem: have a machine convince someone that it is human by means of a test that explicitly disallows inspection of the machine itself. Turing's test conditions don't simply prevent interrogators from seeing the machine; they remove any consideration of how the machine intelligence works from the test.

But given the constraints of the Turing test, this goal is unsatisfactory for many researchers. Their goals are to replicate, or alternatively *understand*, intelligent behavior and intelligence; to build something of substance rather than a façade. Further, the academic world recognizes that intelligence needs a broader definition and that the search for it must look beyond human symbolic intelligence. Robot football and other embodied tasks are now seen as more promising challenges than Turing's test [Brooks 1991; Cohen 2004]. But for game developers, the façade is what counts; it provides a simulation of intelligence to characters in a game world. Believability is more important than truth. Thus the goal of AI in games is generally the same as attempts to beat the Turing test, i.e., to create a believable intelligence by any means necessary.

But for the Turing test – or rather, its derivations– to be applicable to the development of video games, we need to generalize it. Removing the restrictions of the Imitation Game. [Harnad 2000] does just that. Harnad elaborates the Turing test and considers a hierarchy of Turing tests that present goals against which implemented AI systems can be tested and which are not restricted to Turing's conversation-based game. The original "pen-pal" Turing test (a conventionalized version, in which the game can be played over a lifetime rather than in a simple five- minute session) scores low, T2, in Harnad's

hierarchy, which reaches from subtotal or "toy" tests, t1, up to machines which are indistinguishable from humans by any empirical or testable means, T5.

While Harnad focuses his paper on the higher levels of the hierarchy and their philosophical implications, it is the lowest level that is (probably) most applicable to computer games. This is t1, where t stands for toy, not Turing. Toy models of intelligence replicate some subtotal fragments of human functional capacity, from selected arbitrary fragments to collections of self-sufficient modules that represent components of full-scale human capacity.

At the higher levels, being able to pass the Turing test with sentences via a terminal is no longer sufficient – the machine must be able to interact with the world at large. At the lower t1 level, the ability to converse in type is not a requirement at all; any aspect of human functional capacity can be selected for testing. But to pass even a toy test, it is required that human capacities be imitated, not matched or bettered. Beating a grandmaster at chess does *not* mean that a computer has passed the t1 test. Only if the human opponent and other observers are fooled into thinking that a human is playing would the computer pass the t1 test. When Deep Blue defeated the world chess champion Gary Kasparov, the fact that he and other chess experts questioned whether he had been playing a computer raised the issue of whether a computer had passed the Turing test, not the fact that it had defeated him [Krol 1999].

Similarly, the requirement for modern computer games is not unbeatable AI, but believable AI. All we have to do now is determine whether or not some AI is believable. To do this we need some idea of what believable *is* and some means of testing believability. And we can forget about the philosophical arguments over whether or not a machine that fools a human is or is not intelligent; fooling the human is all that we care about.

2. BELIEVABLE AI?

What does it mean to declare that a game AI is believable? The answer may depend quite strongly both on the game being played and the role of AI. To start with, consider two different games set in a historical or contemporary war.

Our first game is a first person shooter (FPS). The players look through the eyes of their characters and control their avatars as they move around a virtual battleground. Other avatars might be controlled by other players or by a computer program. We may judge a computer program a success in any case where a majority of players is unable to tell which of the other avatars are controlled by humans and which by computers. But this is not necessarily the correct goal. Perhaps the challenge should be for the AI avatars to act more convincingly, like the soldiers they represent, acting individually and collectively in a manner that imitates the tactics, training, and combat styles of the real-world combatants. This version is more in keeping with the style of Turing's Imitation Game, challenging the AI to beat a human at some role-playing task. The idea of modifying the Imitation Game in order to have the computer take on different roles is not new, and has been explored previously [Colby 1981; Hayes and Ford 1995]. Which of these goals is correct, should AI be indistinguishable from a human player or should it try to be a better role-player?

If our second game were a strategy game, with the player in the role of a general in command of large numbers of battlefield units, we could pose a similar question**.** Should AI players be tested on whether they play like human players, or whether they use commands that are historically accurate and replicate the behavior of a typical general of the period in question?

The differences between the AI systems that might pass these alternate challenges could be significant; the design of computer games is such that human players often employ tactics unlikely to be used in the real world.

These two examples consider the role of AI in games as only providing an opponent. Yet in current computer games AI is employed to control not just opponent or ally game players but vast arrays of nonplayer characters (NPCs). Consider again the strategy game mentioned above. The AI has to control the behaviors of all the individual soldiers in the game, and to do so in a believable manner. In this case there is no possibility at all that an individual soldier could be played directly by another human – we know that it is played by a machine intelligence. It makes no sense to ask whether this soldier is being played by a human or a machine, so what goal do we have to meet before we can declare the AI for the soldier to be believable? Similarly, other game entities such as wildlife might be expected to demonstrate some amount of intelligence, although clearly not that of a human player.

Hence it is clear that a distinction is to be made is between Player AI, where AI takes on the role of another player, and NPC AI, where it takes on the part of a NPC or a creature that inhabits the game world. We will explore the issues relating to believable AI in both cases, after considering the problem of measuring believability.

3. MEASURING BELIEVABILITY

Judging whether or not some AI behavior could believably be the product of some human (or animal) intelligence is a subjective process. It depends on how an observer perceives the behavior; questions about, and partial solutions to, key problems have been explored by Mac Namee [2004, ch. 6].

The means to determine or compare believability are largely restricted to the use of questionnaires (as in Laird and Duchi [2000], see below), since it depends on the perceptions of people watching or playing games. Mac Namee [2004] notes that the answer to the question of whether something is believable is highly subjective. Scoring believability according to an arbitrary scale is similarly subjective. However, in order to measure believability, some degree of objectivity is desirable; Mac Namee achieves this by asking a different question. Instead of presenting different AIs and asking subjects to rate their believability, he presents just two versions of an AI implementation and asks users which of the two is *more* believable. He additionally requests that subjects comment on any differences that they notice between the two versions.

The results from Mac Namee's own tests highlight that believability is very subjective and can be influenced by cultural effects. In one test, subjects are shown two versions of a virtual bar populated by AI agents. In one version the behavior of the agents is controlled in a deliberate manner because they try to satisfy long-term goals – buy and drink beer, talk to friends, or go to the toilet. In the second version the agents randomly pick new short-term goals every time they complete an action. One Italian subject felt that having agents return to sit at the same table time after time was unrealistic, whereas the other (Irish) subjects mostly felt that this behavior was more believable. Sadly, further tests to determine whether this – or other – results are indeed culturally significant have yet to be carried out; the possibility that such differences exist does appear quite likely, however.

The extent to which cultural differences affect the perception of believability is largely unknown. Current examples of believability testing for games have used rather small numbers of test subjects from narrow locales. If cultural factors do have an impact, there is clearly a need for greater coverage of different user groups.

Another important difference between subjects noted by Mac Namee is between game-playing novices and veterans [Mac Namee 2004, sect. 6.5.2]. To novices, the whole experience of playing a game is so new that they may fail to notice the significant differences between two versions of AI, even where the differences are readily apparent to most veterans.

The commercial value of testing players' perceptions of computer AI was demonstrated in work on the highly successful game, Halo [Butcher and Griesemer 2002]. An interesting result was finding that AI behaviors needed highly exaggerated animations and visible effects in order to be noticeable to players. The implication is that to improve believability it may be required to give characters unrealistically over-emotive reactions and actions. Testing may avoid this effect by having observers rather than players rate the game (again, see the discussion in Laird and Duchi [2000], below); but in game development the aim is to satisfy the needs of the players of a game and not those of watchers.

4. BELIEVABLE COMPUTER PLAYERS

In a very wide range of games, two or more players can play together – either as allies or as opponents, in teams or as individuals. Strategy games in particular allow players to form and break up teams during play. Multiplayer games commonly include AI players, enabling human players to play against or with computer-controlled players. In such a case the applicability of the Turing test and Harnad's Turing hierarchy is clear. We would like AI to play the game in such a human-like way that other human players or observers will find it difficult to distinguish machine from human players. Meeting this challenge would indicate success in passing a t1 test (but not the Turing test), and make for satisfying game-play experiences.

In some cases a human player might not actually be a good judge. As Laird and Duchi [2000] note, human players can be too involved in playing the game, and thus unable to devote much time to observing other players closely. Further, they may be limited to seeing only a fraction of the behavior of the other players – in many games the amount of time that other players are actually in view might be quite small. With limited ability to gather evidence, a human player may be unable to judge the test. However, this constraint may not apply equally to all genres – in strategy games, which a human player still has limited evidence to go on, the slower pace and extended game time may allow players to judge on the basis of opponent strategies and play styles.

4.1 The Soar Quakebot – Almost Human

Accordingly, to evaluate the Soar Quakebot (an AI player for the first-person shooter game Quake II [Laird and van Lent 1999]), Laird and Duchi [2000] had, via video, observers judge the game, rather than have the judges play the game. Five human players with differing degrees of experience played against an expert. The expert played a further 11 games against the Soar Quakebot. For each game in which Quakebot participated, different parameter settings were selected for it. The judges then ranked a nonexpert player in each game for skill and degree of "humanness" (without knowing whether the player is a human or a machine).

Their results show that bots with human-like decision times were rated as more human-like than bots with slower or faster decision times, although no bot scored as high as any human player; and bots with more tactical reasoning were rated more human-like. The judges decided that bots with high aiming skills were too competent to be human, and rated the bots with poorer aim as being more human. While noting that their results

should be considered preliminary, as more extensive testing is required, the judges did highlight a few design principles for creating human-like behavior, as follows:

- Give the AI a human-like reaction and decision time (one twentieth to one tenth of a second).
- Avoid giving the AI superhuman abilities (e.g., overly precise aiming).
- Implement some tactical or strategic reasoning so that the AI is not a purely reactive agent.

4.2 Attempting Human-Like Behavior By Imitating Humans

We also conducted other studies on the believability of AI players [McGlinchey and Livingstone 2004]. The first of these is somewhat simpler than a Quakebot, it involves testing the human-like qualities of AI Pong players. While seemingly trivial, the AI, a self-organizing map trained on human data, is able to replicate the distinct behaviors of different players with some success.

For these tests, a number of Pong games were recorded and played back to test subjects. According to the principles for reducing subjectivity in judging believability noted earlier, subjects were not asked to rate believability. Instead, in each game they were asked which of two players – each might either be human or AI – acted more like a human player. In addition, one game pitched a human versus a hard-coded AI, instead of one trained on human-player data.

When the opposing player hits the ball towards it, the hard-coded AI projects the point the ball will intersect with its bat. After a short delay to simulate human reaction time, the bat is moved gradually towards the point of intersection, with a weight parameter that controls the speed at which the bat is moved to its target position. To make the AI player fallible, the target position for the bat is moved by a random amount, between -20 and +20 pixels from a uniform distribution (the height of the bat is fixed at 64 pixels).

Eight games were recorded and then shown to observers for testing: four games of humans playing against trained AI players; two games between pairs of AI players; one game with two competing human players; and one game with a human player against a hard-coded AI bat.

For each game, the observers were asked whether they thought that the left bat, right bat, both bats, or neither bat was human controlled. A final question asked the observers what evidence, if any, led them to make the decisions they did. Observers were allowed to stop a recording once a decision was reached, and most observers took the opportunity to quit games before concluding.

While the AI had been shown to successfully imitate different play-styles peculiar to different human players [McGlinchey 2003], it was not shown that the imitation was accurate enough to fool human observers. The testing provided some interesting results – results that not only highlight areas for improvement, but demonstrate both the usefulness of believability testing and the care that must be taken in setting questions and analyzing results.

Over the range of returns, it appeared at first that the trained AI performed very well, since it was identified as human about as often as human players were. This appears to confirm that where only very limited aspects of human behavior are observable, the t1 test can be passed easily. A closer look at the results demonstrates more clearly the limits of judging success by the ability to fool observers into believing that some behavior is human.

While one respondent judged 14 out of 16 players correctly, another misclassified 14 out of 16. Many of the subjects varied significantly from chance – either getting most responses correct or most incorrect. This indicates that these observers were able to distinguish between human and machine controllers, even though they made incorrect decisions as to which was which. With this simple system (and very limited domain), the hard-coded AI was also often incorrectly identified. But again a closer look at the data indicates that despite the frequent errors in categorization, the hard-coded AI was visibly distinct from real human players.

With an additional free-text question that asked respondents to explain how they were able to distinguish the two types of players, we were able to check responses to see what aspect of the AI behavior led to the result. The answers show that some of the judges – in particular those who got most identifications right *and* those who got most wrong – noticed that some bats moved with more jerky and sudden movements than others. In most cases these were the AI bats, although some observers thought this behavior was a marker for human control.

4.3 The Believability of AI In Commercial Strategy Games

A different strategy was employed to test the believability of AI players in commercial strategy games, in which we tested the AI players already in the game rather than implementing our own. As noted above, the slower pace of strategy games allows players to act as judges.

Instead of asking if players thought their opponent was human- or computer-controlled, we asked a range of questions to assess whether players believed the computer opponent differed from a human player in any significant way in resources, strategies, or even in ability to cheat. The reasoning behind this is that if players are able to detect distinct patterns in the strategies and tactics employed by a games AI, they may learn to exploit the AI. Indeed, it was the informal observation of such exploitation in games played by the author that provoked the attempt to study this phenomenon.

While this study is currently at an early stage, the results do seem to support the hypothesis that AI in strategy games is exploitable. For example, one common flaw is that AI players, by repeatedly attacking the same place despite continued defeat at that point, become very predictable, allowing players to easily defend while building large armies.

5. BELIEVABLE CHARACTERS

In the last case we presented we were able to compare AI behavior against human behavior, even though the judges *knew* that they were evaluating artificial intelligences. In the case of AI for nonplayer roles, different criteria are required. The AI is not required to act like a player but like an intelligent character or creature within the game world. Compared to the development of opponent AI, research and development on AI support-characters has been more limited [MacNamee 2004, p.123].

It is also worth noting that the characters do not even need to be human. Harnad [2000, p. 432] notes that a Turing hierarchy can be posited for any creature, and so there can be a t1 toy model for any creature we might wish to include in a game – mouse, monkey, or even mollusk.

If we capture enough of the behavior of a creature and replicate it in the virtual world so that an observer agrees that it does indeed act and behave like a real creature, then we could argue that it satisfies the t1 test. After letting observers watch video sequences of games, we present them with questionnaires to determine their opinion as to how realistic the observed behavior was. In games set in other worlds populated with imagined

creatures, we can ask observers the hypothetical "does the behavior of the orcs in the game match the expected behavior?"

While creating believable animal behavior may be quite different from creating believable human behavior, the evaluation method is very similar**,** that is, via surveys and questionnaires. The tests carried out by Mac Namee [2004] on believable characters are good examples for testing believability for character AI. Although some people seem adept at noticing unrealistic character behavior, such behavior is readily accepted by others. And, to repeat, in their work on AI testing on Halo, Butcher and Griesemer [2002] noted that not all players are aware of what reasoning (if any) lies behind a character's actions, even when, in an attempt to make the AI's behavior as transparent as possible, the actions are somewhat exaggerated.

5.1 Unbelievable Characters

If players don't notice AI failures, then why bother to test believability at all? To remind ourselves of the value of believability, it is worth considering a typical example of how game characters might fall short of realistic behavior, which in an extreme case may cause players to suspend the suspension of disbelief (a problem noted by Wetzel [2004]). The following example is from the game Fallout: Tactics, but the nature of the failure (NPC guards fail to react to obvious signs that something is amiss) is common to a large number of games [Wetzel 2004].

At one point in the game, two enemy guards (they are not each other's enemies) patrol a wall and must be passed. The two guards both follow the same path, starting at opposite ends and passing each other in the middle. It is possible to lay a land mine so that the first guard walks into it and dies messily and loudly. The other guard, perhaps 100 yards away, continues his patrol, oblivious. As he crosses the passing point where he normally meets his compatriot he fails to notice his compatriot's absence. He continues his patrol, walking through the bloody remains of his companion and back again without reaction. It is possible to set a second land mine at the blood-stained wall in the same position as the first, and the second guard will walk straight into it, his doom sealed by failure to pick up on any of the obvious clues.

This behavior will certainly cause the suspension of disbelief to disappear, as this behavior is clearly unrealistic. The importance of the suspension of disbelief is emphasized in a recent paper by Sweetser and Wyreth [2005] which, although it does not consider the significance of believable AI, highlights the degree to which a player's enjoyment of a game depends on his immersion in it. The severity of the problem is made clear by Wetzel [2004], who describes an attempt to build a list of the most prevalent and common AI failures in games. He finds a disturbing number of examples of bad AI which, with a little effort, seem easily avoidable.

6. THE EMBEDDED TURING TEST

We have overlooked, until now, the possibility of embedding the T2 Turing test in a game. Many multiplayer games allow players to type and send messages to other players. With no limits on what may be typed, Turing's Imitation Game can easily be embedded in such games. If the AI in an embedded Turing test is to play the game while chatting, then it is a slightly extended T2 test – although playing the game in a convincingly human manner may be a much simpler challenge than chatting in a convincingly human manner.

In many multiplayer action games the chat between players is actually quite limited in range, often consisting of little more than taunts and abuse. To produce this would present little challenge for current AI techniques. Other types of games, e.g., online role-

playing games or strategy games, often have more extended chat between players. But conversation within the confines of such games may still be significantly more limited in topic and depth than in a conventional Turing test. Thus making embedded Turing tests in many ways easier than normal – although this has yet to be tested formally.

While Turing does not clearly state whether the interrogator is aware that one of the contestants might not be human, most often the conventionalized Turing test appears to have this awareness as a condition of the test. There are reported cases where people have been fooled by chatbots (conversational artificial intelligence programs), which have for many years been used in text-based MUDs (Multi-User Dungeons) and chat programs. Reportedly, individuals have chatted to bots for significant amounts of time without realizing that they were chatting with machines [Humphrys 1995]. In such cases it appears that the individuals who were fooled by bots were unaware of the possibility that they could be chatting with a machine. If the people involved in these conversations had been aware of this possibility, then it is questionable whether they would have been fooled at all.

7. CRITERIA FOR BELIEVABLE GAME AI

We previously declared that believability is a nebulous concept, but to make use of the ideas discussed here some guidance on how to apply this knowledge is clearly required. To this end, we present a set of criteria that might be usefully applied in deciding whether or not a particular game AI is believable.

For Player AI, which tries to play the game as a human player would (as either opponent or ally), Turing's own test criteria are clearly applicable but not necessarily particularly helpful. We have seen (above) examples of how human testers can play against, or simply observe, game AI in action before pronouncing judgment on the "humanness" of its behavior. In a blind test that mixes AI and human players this can certainly give some indication of overall success, although with some limitations, as shown by the Pong testing.

A key problem with Turing's criterion is that, as noted in the introduction, it is an all-or-nothing test: the game AI will either pass or fail. A set of criteria that outlined behaviors required for human-like play would be more useful. With such criteria it would be possible to test against different aspects of believable play more easily, and allow developers to focus on improving areas where the AI is weak, as well as provide a check-list against which to test the AI in the first place. The more detailed criteria would also allow the possibility for conducting more evaluation by expert review. This would reduce, but might not eliminate, the need for expensive evaluation by gamers.

Another problem with Turing's criterion is that it simply does not apply to NPC AI, where it is generally known by players *a priori* that the characters are computer-controlled. Here the more detailed criteria provide a means for evaluating behavior without having to consider whether the controller behind the behavior is human or not.

A possible set of believability criteria is presented in Table I; these were distilled from the discussions above, from the findings of Laird and Duchi [2000], and from Wetzel's attempt to provide a list of the most prevalent AI flaws in games [Wetzel 2004]. In the distillation, we looked for common threads in observed AI failures to produce a manageable set of criteria. We divided the AI behaviors into three broad categories, i.e., Plan, Act, and React. The criteria remain somewhat open; and for each criterion the precise behaviors required will naturally vary according to the nature of the game. The criteria do not distinguish between games and simulations or the resultant difference in AI behavior, so developers will need to be aware of the audience (see Section 2).

Table I. The PAR AI Believability Criteria

	AI should…
Plan[1]	(1.1) demonstrate some degree of strategic/tactical planning (1.2) be able to coordinate actions with player/other AI (1.3) not repeatedly attempt a previous, failed, plan or action
Act[2]	(2.1) act with human-like reaction times and abilities
React[3]	(3.1) react to players' presence and actions appropriately (3.2) react to changes in their local environment (3.3) react to presence of foes and allies
Notable exceptions	1 Might not apply where design/plot calls for impulsive or stupid characters, nor for animals 2 Might not apply where design/plot calls for characters with significantly superior or inferior abilities 3 Might not apply where game-design/plot call for characters with limited awareness

Plan

Any game AI will need to make decisions, so it is important that some strategy or tactic informs these decisions [Laird and Duchi 2000; Mac Namee 2004]. AI certainly should avoid making obviously wrong decisions, such as throwing grenades at adjacent opponents and consequently self-destructing [Wetzel 2004].

While on the surface the need for coordinated behavior might apply to only a subset of games, it is actually a very common requirement. For instance, simply moving a small number of AI characters through a narrow area without having them block and impede each other has tripped up many games [Wetzel 2004]. In action and strategy games there is often a need for groups to coordinate attacks, concentrate forces, and avoid friendly fire, which again is often beyond current AI implementations [Wetzel 2004]. Strategy games are particularly, but not exclusively, prone to creating AI opponents that repeatedly attack the same positions, even when such attacks fail repeatedly (see Section 4.3).

Act

When a decision has been taken, the AI needs to appear to have natural reaction times, and to be able to act with a reasonable degree of skill [Laird and Duchi 2000; McGlinchey 2003].

React

Wetzel [2004] provides a number of AI examples that simply fail to react to player actions, to their environments, and to other game entities. AI that perform well according to some criteria do not always do so to others, hence the presentation of three distinct criteria here. For example, a patrolling guard who does not react to the sudden appearance of blood stains fails to notice a change in environment (see discussion of Fallout: Tactics, above). Such a failure may co-exist with a reasonable degree of believable behavior in reacting to the player and other game entities.

Believability Is Not the Only Goal

This is an appropriate point at which to note that believability is not the only goal of game AI. The primary goal, helped by believable AI, is an entertaining game [Charles 2003]. In order to help evaluate player enjoyment, an expansive set of criteria, the

GameFlow criteria, were recently proposed in the paper by Sweetser and Wyreth [2005]. While that paper does not include a discussion of the effect of AI, AI clearly has an effect on the GameFlow criteria for immersion and challenge, and possible effects on other criteria such as control and player skills. Not only should AI be believable but it should help provide a suitably, not overly, challenging experience.

8. CONCLUSIONS

Game AI that is capable of fooling us into thinking it is another human player cannot be considered as having passed the Turing test because it was not playing either Turing's or the conventional version of the Imitation Game. So it is ironic that while Turing's game has been largely rejected by mainstream AI research, a machine playing Turing's game shares the same fundamental goal with video game AI: to fool people into thinking that there is a genuine (human) intelligence at work where there is only a computer program. The idea that game AI should be believable is a nebulous goal that requires careful definition each time it is applied, as the behaviors that game AI should exhibit will vary strongly, depending on what it is supposed to imitate. The different types of AI are primarily distinguished according to two goals: to emulate the behavior of players and to create lifelike characters. This distinction gives clearer goals for AI behavior and extends believability testing to cases where testers know in advance that they are scrutinizing machine intelligence.

A common goal for game AI developers, one that is more tightly defined than believability, is likely to remain elusive. The tasks required of AI vary tremendously, and many games have only limited and tenuous dependence on reality (alien wars, spaceships that decelerate in the absence of propulsion, anthropomorphic animals, and general magic and mayhem). Further, as AI becomes more sophisticated and believable, so too are players' expectations also evolving over time [Charles 2003]. This, allied to the discovery that different players already respond differently to game AI according to experience and cultural factors, means that evaluating AI is likely to remain problematic.

It is necessary to recognize that people who play few games have such limited experience of game AI that we cannot expect them to sensibly evaluate different versions of an AI. By explicitly adding evaluation of game AI to the quality assurance process (by asking people who play games of the type being tested) developers can identify the strengths and weaknesses of their AI. The feedback from novices, while interesting and potentially useful, is less useful for evaluating the AI of a game. As Wetzel [2004] notes: the first step in building better AI is to document its failures – and, as we have seen, experienced game players are more likely to recognize failures when they see them.

Accordingly, we argue that the opinions and experiences of players need to be recorded and documented as an essential part of the process of improving the quality of game AI. In this we are in agreement with Sweetser and Drenna [2003] and Purdey and Thornton [2004], who argue more generally for recording the opinions and feelings of players as important for improving future game design; such a process is expensive, however.

We close this article with a presentation of criteria that might be used for evaluating the believability of game AI. The criteria help break down the overall goal into three broad categories, with some further break-down of requirements within categories. The set of criteria provides a basic framework that enables evaluation of believability in more detail than a simple binary result, and which we hope will provide a meaningful basis for expert review by developers. To improve the quality of game AI and relying on play-

testing alone to find believability issues is likely to provide feedback too late to be of use. Developers themselves need to consciously, and very deliberately, look for flaws.

9. AFTERWORD: THE IN-GAME IMITATION OF THE IMITATION GAME

Hayes and Ford [1995] conclude their critical look at the negative effect of the Turing test on AI research with the argument that the Turing test works better as a means to provoke thought about the nature of the "human condition" than it does as a useful measure of machine intelligence: "We suspect that Turing chose this topic because he wanted the test to be about what it really means to be human" [Hayes and Ford 1995].

The idea of thinking about machine intelligence as a means to framing questions about human intelligence and nature was used by the cult science-fiction author Philip K Dick. In *Do Androids Dream of Electric Sheep?* [Dick 1968], the only way to determine if someone is an android – a machine that has human-like intelligence and is visually indistinguishable from humans – is by means of a Turing test-inspired device: the Voight-Kampff test.

As a subject is interrogated with a barrage of questions, sensors help an interrogator observe and detect minute biological responses. The message of *Do Androids Dream of Electric Sheep?* is that there *is* more to being human than being intelligent: there is emotion and empathy. It is the failure of the artificially constructed humans to respond empathically with the associated biological responses that separates them from real people. We could also compare the use of questionnaires in the book to detect sociopath and psychopath-like androids with the modern use of questionnaires to detect human psychopaths [Hare 1993], (but that is a topic for another paper and another journal).

Do Androids Dream of Electric Sheep? was later made into a film called BladeRunner [Dick et al. 1982]. Later still, a computer game based on the film was released [Westwood 1998]. While playing this game, a single human player can direct their character as it conducts Voight-Kampff tests on other characters in the game.

In the game the player acts as interrogator in an imitation Imitation Game, in an attempt to determine which other characters are machines. But this is a video game, all the characters are machines.

END NOTES

1. More information on the annual Loebner Prize competition can be found at: http://www.loebner.net/Prizef/loebner-prize.html. It is interesting to note that despite the difficulty still facing competitors in successfully fooling a panel of judges in a Turing test imitation game, the requirements for winning the Loebner gold medal prize now requires the winner's work to respond in a human-like way to images and videos. Clearly, it will be some time before a gold medal is awarded.

2. Krol [1999] notes that the highest-ranking computer scored points close to that of the lowest-ranking humans in the 1998 competition. She [Krol] also notes that some of the transcripts reveal that some human confederates were trying to fool judges into thinking that they were computers. Despite this, no program managed to rank higher than any human. Just two years later the judges' predictions were 91% correct after just 5 minutes [Akman and Blackburn 2000].

ACKNOWLEDGMENTS

The author would like to thank Bronwen, Darryl Charles, and the anonymous reviewer for their constructive comments, and the Carnegie Trust for the Universities of Scotland for supporting this work.

BIBLIOGRAPHY

AKMAN, V. AND BLACKBURN P. 2000. Editorial: Alan Turing and artificial intelligence. *J. Logic, Language and Information 9* (2000), 391-395.

BROOKS, R. A. 1991. Intelligence without representation. *Artificial Intelligence J. 47* (1991), 139–159.

BUTCHER, C. AND GRIESEMER, J. 2002. The illusion of intelligence: The integration of AI and level design in Halo. Presented at the *Game Developers Conference* (San Jose, CA, March 21-23, 2002).

CHARLES, D. 2003. Enhancing gameplay: Challenges for artificial intelligence in digital games. In *Level Up: Digital Games Research Conference Proceedings* (Utrecht, Nov. 4-6).

COHEN, P. 2004. If not the Turing test, then what? In *Proceedings of the Nineteenth National Conference on Artificial Intelligence* (San Jose, CA, July 25-29, 2004). American Association of Artificial Intelligence.

COLBY, K. M. 1981. Modeling a paranoid mind. *Behavioral and Brain Sciences 4* (1981), 515-560.

DICK, P. K. 1968. *Do Androids Dream of Electric Sheep?*

DICK, P. K., FANCHER, H., AND PEOPLES, D. 1982. Blade Runner. R. Scott (Director).

HARE, R. D. 1993. *Without Conscience: The Disturbing World of the Psychopaths Among Us*. Simon & Schuster, New York.

HARNAD, S. 2000. Minds, machines and Turing. *J. Logic, Language and Information 9* (2000), 425-445.

HAYES, P. J. AND FORD, K. M. 1995. Turing test considered harmful. In *Proceedings of International Joint Conference on Artificial Intelligence* (IJCAI-95, Montreal, Quebec).

HUMPHRYS, M. 1995. How my program passed the Turing test. http://computing.dcu.ie/~humphrys/eliza.html. Online Mar. 1, 2005.

KROL, M. 1999. Have we witnessed a real life Turing test? *IEEE Computer* (March, 1999), 27-30.

LAIRD, J. E. AND DUCHI, J. C. 2000. Creating human-like synthetic characters with multiple skill levels: A case study using the Soar Quakebot. In *Proceedings of the AAAI 2000 Fall Symposium: Simulating Human Agents* (North Falmouth, MA, Nov. 3-5, 2000).

LAIRD, J. E. AND VAN LENT, M. 1999. Developing an artificial intelligence engine. In *Proceedings of the Game Developers Conference* (San Jose, CA, Nov. 3-5, 1999).

MAC NAMEE, B. 2004. Proactive persistent agents: Using situational intelligence to create support characters in character-centric computer games. Ph.D. dissertation, Dept. of Computer Science. University of Dublin, Dublin, Ireland

MCGLINCHEY, S. 2003. Learning of AI players from game observation data. In *Proceedings of the Game-On 2003, 4th International Conference on Intelligent Games and Simulation* (London).

MCGLINCHEY, S. AND LIVINGSTONE, D. 2004. What believability testing can tell us. In *CGAIDE, Proceedings of the Conference on Game AI, Design and Education* (Microsoft Campus, Redding, WA).

PURDY, J. H. AND THORNTON, J. S. 2004. What effects should a game have on brain & body to be successful? In *Proceedings of the Game Developers Conference Europe* (London, Aug. 31-Sept. 3, 2004).

SEARLE, J. R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences 3* (1980), 417-424.

SWEETSER, P. AND DRENNA, P. 2003. User-centred design in games. Presented at the *Australian Game Developers Conference Academic Summit* (Melbourne, Nov. 20, 2003).

SWEETSER, P. AND WYETH, P. 2005. GameFlow: A model for evaluating player enjoyment in games. *ACM Computers In Entertainment 3*, 3 (2005). Online only.

TURING, A. M. 1950. Computing machinery and intelligence. *Mind LIX*, 236 (1950), 433-460.

WESTWOOD. 1998. Blade Runner (PC CD-Rom).

WETZEL, B. 2004. Step one: Document the problem. In *Proceedings of the AAAI Workshop on Challenges in Game AI* (San Jose, CA, July 25-26, 2004).

Received May 2005; accepted October 2005